APPENDIX M

STATISTICAL METHODS FOR COMPARING SAMPLES: SPATIAL AND TEMPORAL CONSIDERATIONS

APPENDIX M

STATISTICAL METHODS FOR COMPARING SAMPLES: SPATIAL AND TEMPORAL CONSIDERATIONS

The primary objective of Tier 2 intensive studies is to assess the magnitude and geographic extent of contamination in selected target species by determining whether the mean contaminant concentration exceeds the screening value (SV) for any target analyte. Secondary objectives of intensive studies may include defining the geographical region where fish contaminant concentrations exceed screening values (SVs), identifying geographic distribution of contaminant concentrations, and, in conjunction with historical or future data collection, assessing changes in fish contaminant concentrations over time. This appendix discusses some of the statistical methods that may be used to compare fish contaminant levels measured at different locations or over time.

The recommended statistical approach for comparing replicated contaminant measurements between two or more groups is outlined below and in Figure M-1. For each type of test, several options are provided, each of which may be appropriate in specific cases. State staff should consult a statistician as to the specific statistical tests to use for a particular data set.

Statistical tests of significant differences between means (or other measures of central tendency) can be divided into parametric and nonparametric types. Parametric tests assume that the contaminant concentrations in the population being sampled are normally distributed and that the population variances in the groups being tested are not significantly different from each other (Gilbert, 1987). If either of these assumptions is violated, a nonparametric test may be more appropriate. However, nonparametric tests should be used only when necessary because the power of parametric tests generally is greater than the power of nonparametric tests when the assumptions of the parametric test have been met (Sokal and Rohlf, 1981).

Because the populations of many environmental measurements are not normally distributed, logarithmic transformation is often performed on the sampled data (Gilbert, 1987). However, transformation may not be appropriate in all cases. If the data are sampled from a population that is normally distributed, then there is no need for transformation (Figure M-1).

If the assumptions of normality and equality of variance are met, parametric tests of significant differences between means, such as the one-way Analysis of Variance (ANOVA) and the *t*-test, should be performed. If three or more groups are compared using the ANOVA that results in a significant difference, the

Figure M-1. Statistical approach to testing for significant differences between different groups of contaminant monitoring data.

difference in mean concentrations between two group means can be further investigated using a multiple comparison test (Figure M-1). These tests indicate which specific means are significantly different from each other, rather than just indicating that one or more means are different, as the ANOVA does.

If the underlying assumptions for parametric testing are not met, nonparametric tests of significance can be employed. Nonparametric tests of significant differences in central tendencies are often performed on transformed data, that is, the ranks. Multiple comparison tests comparable to those used for parametric data sets are not available for nonparametric data sets. For data sets including three or more groups, a series of two-sample tests can be performed that can yield similar information to that derived from multiple comparison tests.

Because the concentrations of contaminants, particularly nonpolar organics, are often correlated with the percentage of lipid in a tissue sample (see Section 8.1.2), contaminant data are often normalized to the lipid concentration before statistical analyses are performed. This procedure can, in some instances, improve the power of the statistical tests. States wishing to examine the relationship between contaminant concentrations and percentage of lipid should refer to Hebert and Keenleyside (1995) for a discussion of the possible statistical approaches.

Intensive studies may include the collection of fish contaminant data from several locations within a region of interest or for multiple time periods (e.g., seasons or years) from a single location, or a combination of both. Data from intensive studies such as these may be used to perform spatial (i.e., between stations) or temporal (i.e., over time) analyses. It should be noted that these types of analyses, if performed, are performed in addition to the statistical comparisons of mean target analyte concentrations with SVs described in Section 6.1.2.7. It is only the latter type of comparison that should be used to make decisions regarding the necessity of performing risk assessments and the issuance of fish consumption advisories. Spatial and temporal comparisons of contaminant data, however, may yield important information about the variability of target analyte concentrations in specific populations of a particular target species.

M.1 SPATIAL COMPARISON OF STATIONS

Intensive studies also may involve the collection of contaminant data from multiple stations within a waterbody of interest. The stations could be located in different lakes within a single drainage basin, upstream and downstream of a point source of concern along a single river, or randomly located within a single waterbody if an estimate of random spatial variability is desired. The use of an example will serve to illustrate how a spatial analysis of contaminant data might be performed. In this example, a State has determined from a screening study on a river that cadmium is present in a target species at 20 ppm, which is two times the SV of 10 ppm (see Table 5-2). An intensive survey was undertaken in which eight samples were collected from three locations on the river of potential concern and analyzed for cadmium. The results of the analyses for

each location and the statistical comparisons between the three groups are presented in Table M-1.

The mean cadmium concentration at each of three locations was more than twice the SV of 10 ppm (Table M-1). The most important statistical test, as indicated in Section 6.1.2.7, is a comparison of the mean target analyte concentration for each location with the appropriate SV for that target analyte using a *t*-test. These tests must be performed before any analysis of spatial trends is performed. The results of the *t*-tests indicate that each of the three mean tissue concentrations is significantly greater than the SV (Table M-1). By itself, these results indicate that a risk assessment is warranted.

Table M-1. Hypothetical Cadmium Concentrations (ppm) in Target Species A at Three River Locations

Replicate samples	Station 1	Station 2	Station 3
1	20	28	33
2	18	27	30
3	25	34	30
4	22	28	28
5	21	30	20
6	22	29	39
7	23	30	31
8	21	29	30
Mean	21.5	29.4	31.3
Standard deviation	2.07	2.13	3.45
p-Value for *t*-test with SV	<0.001	<0.001	<0.001
p-Value for W test	0.97	0.83	0.78
p-Value for Levene's test		0.52	
p-Value for ANOVA		<0.0001	
p-Value for Duncan's-1 vs. 2		<0.0001	
p-Value for Duncan's-1 vs. 3		>0.0001	
p-Value for Duncan's-2 vs. 3		0.17	

A general statistical flowchart for comparing contaminant concentration data from several stations to each other is presented in Figure M-1. The cadmium data in Table M-1 may be additionally analyzed using the tests in Figure M-1. All of the statistical tests in Figure M-1 can be performed using commercial statistical software packages. By performing a spatial analysis of the data, the details of the risk assessment might be further refined. For example, one component of a fish advisory is often the establishment of risk-based consumption limits (see Volume II of this series). In order to calculate these limits, an estimate of the contaminant concentration in the target species must be available. In the example shown in Table M-1, there are three estimates of cadmium concentration. A spatial analysis of these data can help to identify which of the concentrations (if any) to use in establishing risk-based consumption limits.

The initial steps in the flowchart on Figure M-1 are to determine whether parametric or nonparametric statistical tests should be used. The first step is to test whether each of the three groups of data are from populations that are normally distributed. Three tests that may be used for this purpose are the Kolmogorov-Smirnov test for normality (Massey, 1951), Shapiro and Wilk's W test (Shapiro et al., 1968; Royston, 1982), and Lilliefors' test (Lilliefors, 1967). The results for the W test on each of the three groups of data indicate that each group was sampled from populations that are normally distributed (Table M-1). The next step is to test for homogeneity of variances between the three groups. Three tests that may be used for this purpose are Levene's test (Milliken and Johnson, 1984), the Hartley F-max test (Sokal and Rohlf, 1981), and the Cochran C test (Winer, 1962). The result of Levene's test indicates that the variances of the three groups of data are not significantly different from each other (Table M-1). These test results mean that parametric statistics (the left side of Figure M-1) are appropriate for this dataset.

An appropriate parametric test to perform to determine whether the three mean cadmium concentrations are significantly different from each other is a 1-way ANOVA. The result of this test indicates that the three means are significantly different (Table M-1). What this result does not show, however, is whether each mean concentration is significantly different from both of the other mean concentrations. For this answer, multiple comparison tests can be used to perform all possible pairwise comparisons between each mean.

Three tests that can be used to perform a multiple comparison are the Newman-Keul test (Sokal and Rohlf, 1981), Duncan's Multiple Range test (Hays, 1988; Milliken and Johnson, 1984), and the Tukey Honest Significant Difference test (Hays, 1988; Milliken and Johnson, 1984). Three pairwise comparisons are possible between three means (1 vs. 2, 1 vs. 3, and 2 vs. 3). The results of Duncan's Multiple Range test indicate that the mean concentration at station 1 (21.5 ppm) is significantly lower than the mean concentrations at both station 2 (29.4 ppm) and station 3 (31.3 ppm), which in turn are not significantly different from each other. Therefore, to be most conservative (i.e., protective), the State could use the mean of the 16 replicate samples from stations 2 and 3 to calculate risk-based consumption limits. In this example, use of the concentra-

tion from any single station would not truly represent the potential contaminant exposure to fish consumers in the waterbody of concern.

M.2 TEMPORAL COMPARISON OF STATIONS

Both screening and intensive studies are often repeated over time to ensure that public health is adequately protected. By examining monitoring data from several time periods from a single site, it may be possible to detect trends in contaminant concentrations in fish tissues. Trend analysis data should never be used to conduct risk assessments. Procedures for conducting risk assessments are adequately covered elsewhere in this document (see Section 6.1.2.7). Trend analysis may, however, be useful for monitoring the effects of various environmental changes or policies on the contaminant concentrations in the target species. For example, a State may have issued a fish advisory for a contaminant for which the source is known or suspected. Source control for this contaminant is the obvious solution to the environmental problem. An evaluation of the effectiveness of the source control may be made easier by trend analysis. The State would still need to perform statistical calculations comparing data from each sampling site to the SV, but trend analysis could yield valuable information about the success of remediation efforts even if the fish advisory remained in place because of SV exceedances.

Trend analysis can be performed using the statistical framework outlined in Figure M-1, but complexities in pollution data collected over time may make this approach unsuitable in some instances. The types of complexities for which other statistical approaches might be warranted can be divided into four groups: (1) changes in sampling and/or analysis procedures, (2) seasonality, and (3) correlated data (Gilbert, 1987). Each of these subjects is discussed briefly here.

Changes in the designation of an analytical laboratory to perform analyses or changes in sampling and/or analytical procedures are not uncommon in long-term monitoring programs. These changes may result in shifts in the mean or variance of the measured values, which could be incorrectly attributed to natural or manmade changes in the processes generating the pollution (Gilbert, 1987). Ideally, when changes occur in the methods used by the monitoring program, comparative studies should be performed to estimate the magnitude of these changes.

Seasonality may introduce variability that masks any underlying long-term trend. Statistically, this problem can be alleviated by removing the cycle before applying tests or by using tests unaffected by cycles (Gilbert, 1987). Such tests will not be discussed here. States interested in performing temporal analyses with data for which a seasonal effect is hypothesized should consult the nonparametric test developed by Sen (1968) or the seasonal Kendall test (Hirsch et al., 1982).

Measurements of contaminant concentrations taken over relatively short periods of time are likely to be positively correlated. Most statistical tests, however,

including those in Figure M-1, require uncorrelated data. Gilbert (1987) discusses several methods for performing the required analyses in these cases.

Temporal trends in contaminant concentrations may be detected by regression analyses, whereby the hypothesis is tested that concentrations are not changing in a predictable fashion (usually linear) over time. If the hypothesis is rejected, a trend may be inferred. States interested in performing regression analyses should consult statistics textbooks such as Gilbert (1987) or Snedecor and Cochran (1980).

M.3 REFERENCES

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold Company, New York, NY. 320 pp.

Hays, W.L. 1988. *Statistics*. Fourth Edition. CBS College Publishing, New York, NY.

Hebert, C.E., and K.A. Keenleyside. 1995. To normalize or not to normalize? Fat is the question. *Environmental Toxicology and Chemistry* 14(5):801-807.

Hirsch, R.M., J.R. Slack, and R.A. Smith. 1982. Techniques of trend analysis for monthly water quality data. *Water Resources Research* 18:107-121.

Lilliefors, H.W. 1967. The Kolmogorov-Smirnov test for normality with mean and variance unknown. *J. Amer. Stat. Assoc.* 62:399-402.

Massey, F.J., Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *J. Amer. Stat. Assoc.* 46:68-78.

Milliken, G.A., and D.E. Johnson. 1984. *Analysis of Messy Data: Volume 1. Designed Experiments*. Van Nostrand Reinhold Company, New York, NY.

Royston, J.P. 1982. An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics* 31:115-124.

Sen, P.K. 1968. On a class of aligned rank order tests in two-way layouts. *Annals of Mathematical Statistics* 39:1115-1124.

Shapiro, S.S., M.B. Wilk, and H.J. Chen. 1968. A comparative study of various tests of normality. *J. Amer. Stat. Assoc.* 63:1343-1372.

Snedecor, G.W., and W.G. Cochran. 1980. *Statistical Methods*. 7th edition. Iowa State University Press, Ames, Iowa.

Sokal, R.R., and F.J. Rohlf. 1981. *Biometry. The Principles and Practice of Statistics in Biological Research*. Second Edition. W.H. Freeman and Company, New York, NY. 859 pp.

Winer, B.J. 1962. *Statistical Principles in Experimental Design*. McGraw-Hill, New York, NY.